SemEval-2010 Task 7: Argument Selection and Coercion

James Pustejovsky and **Anna Rumshisky** and **Alex Plotnick**
Dept. of Computer Science
Brandeis University
Waltham, MA, USA

Elisabetta Jezek
Dept. of Linguistics
University of Pavia
Pavia, Italy

Olga Batiukova
Dept. of Humanities
Carlos III University of Madrid
Madrid, Spain

Valeria Quochi
ILC-CNR
Pisa, Italy

Abstract

We describe the *Argument Selection and Coercion* task for the SemEval-2010 evaluation exercise. This task involves characterizing the type of compositional operation that exists between a predicate and the arguments it selects. Specifically, the goal is to identify whether the type that a verb selects is satisfied directly by the argument, or whether the argument must change type to satisfy the verb typing. We discuss the problem in detail, describe the data preparation for the task, and analyze the results of the submissions.

1 Introduction

In recent years, a number of annotation schemes that encode semantic information have been developed and used to produce data sets for training machine learning algorithms. Semantic markup schemes that have focused on annotating entity types and, more generally, word senses, have been extended to include semantic relationships between sentence elements, such as the semantic role (or label) assigned to the argument by the predicate (Palmer et al., 2005; Ruppenhofer et al., 2006; Kipper, 2005; Burchardt et al., 2006; Subirats, 2004).

In this task, we take this one step further and attempt to capture the "compositional history" of the argument selection relative to the predicate. In particular, this task attempts to identify the operations of type adjustment induced by a predicate over its arguments when they do not match its selectional properties. The task is defined as follows: for each argument of a predicate, identify whether the entity in that argument position satisfies the type expected by the predicate. If not, then identify how the entity in that position satisfies the typing expected by the predicate; that is, identify the source and target types in a type-shifting or *coercion* operation.

Consider the example below, where the verb *report* normally selects for a human in subject position, as in (1a). Notice, however, that through a metonymic interpretation, this constraint can be violated, as demonstrated in (1b).

(1) a. John reported in late from Washington.
  b. Washington reported in late.

Neither the surface annotation of entity extents and types nor assigning semantic roles associated with the predicate would reflect in this case a crucial point: namely, that in order for the typing requirements of the predicate to be satisfied, a *type coercion* or a *metonymy* (Hobbs et al., 1993; Pustejovsky, 1991; Nunberg, 1979; Egg, 2005) has taken place.

The SemEval Metonymy task (Markert and Nissim, 2007) was a good attempt to annotate such metonymic relations over a larger data set. This task involved two types with their metonymic variants: *categories-for-locations* (e.g., place-for-people) and *categories-for-organizations* (e.g., organization-for-members). One of the limitations of this approach, however, is that while appropriate for these specialized metonymy relations, the annotation specification and resulting corpus are not an informative guide for extending the annotation of argument selection more broadly.

In fact, the metonymy example in (1) is an instance of a much more pervasive phenomenon of type shifting and coercion in argument selection. For example, in (2) below, the sense annotation for the verb *enjoy* should arguably assign similar values to both (2a) and (2b).

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 27–32,
Uppsala, Sweden, 15-16 July 2010.

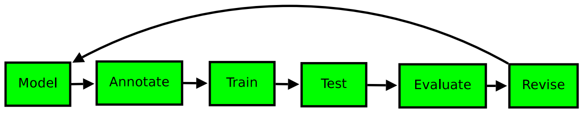

Figure 1: The MATTER Methodology

(2) a. Mary enjoyed drinking her beer.
    b. Mary enjoyed her beer.

The consequence of this is that under current sense and role annotation strategies, the mapping to a syntactic realization for a given sense is made more complex, and is in fact perplexing for a clustering or learning algorithm operating over subcategorization types for the verb.

2 Methodology of Annotation

Before introducing the specifics of the argument selection and coercion task, we will briefly review our assumptions regarding the role of annotation in computational linguistic systems.

We assume that the features we use for encoding a specific linguistic phenomenon are rich enough to capture the desired behavior. These linguistic descriptions are typically distilled from extensive theoretical modeling of the phenomenon. The descriptions in turn form the basis for the annotation values of the specification language, which are themselves the features used in a development cycle for training and testing a labeling algorithm over a text. Finally, based on an analysis and evaluation of the performance of a system, the model of the phenomenon may be revised.

We call this cycle of development the MATTER methodology (Fig. 1):

Model: Structural descriptions provide theoretically informed attributes derived from empirical observations over the data;

Annotate: Annotation scheme assumes a feature set that encodes specific structural descriptions and properties of the input data;

Train: Algorithm is trained over a corpus annotated with the target feature set;

Test: Algorithm is tested against held-out data;

Evaluate: Standardized evaluation of results;

Revise: Revisit the model, annotation specification, or algorithm, in order to make the annotation more robust and reliable.

Some of the current and completed annotation efforts that have undergone such a development cycle include PropBank (Palmer et al., 2005), NomBank (Meyers et al., 2004), and TimeBank (Pustejovsky et al., 2005).

3 Task Description

The argument selection and coercion (ASC) task involves identifying the selectional mechanism used by the predicate over a particular argument.[1] For the purposes of this task, the possible relations between the predicate and a given argument are restricted to *selection* and *coercion*. In *selection*, the argument NP satisfies the typing requirements of the predicate, as in (3):

(3) a. The spokesman denied the statement (PROPOSITION).
    b. The child threw the stone (PHYSICAL OBJECT).
    c. The audience didn't believe the rumor (PROPOSITION).

Coercion occurs when a type-shifting operation must be performed on the complement NP in order to satisfy selectional requirements of the predicate, as in (4). Note that coercion operations may apply to any argument position in a sentence, including the subject, as seen in (4b). Coercion can also be seen as an object of a proposition, as in (4c).

(4) a. The president denied the attack (EVENT → PROPOSITION).
    b. The White House (LOCATION → HUMAN) denied this statement.
    c. The Boston office called with an update (EVENT → INFO).

In order to determine whether type-shifting has taken place, the classification task must then involve (1) identifying the verb sense and the associated syntactic frame, (2) identifying selectional requirements imposed by that verb sense on the target argument, and (3) identifying the semantic type of the target argument.

4 Resources and Corpus Development

We prepared the data for this task in two phases: the *data set construction phase* and the *annotation phase* (see Fig. 2). The first phase consisted of (1) selecting the target verbs to be annotated and compiling a sense inventory for each target, and (2) data extraction and preprocessing. The prepared data was then loaded into the annotation interface. During the annotation phase, the annotation judgments were entered into the database, and an adjudicator resolved disagreements. The resulting database was then exported in an XML format.

[1]This task is part of a larger effort to annotate text with compositional operations (Pustejovsky et al., 2009).

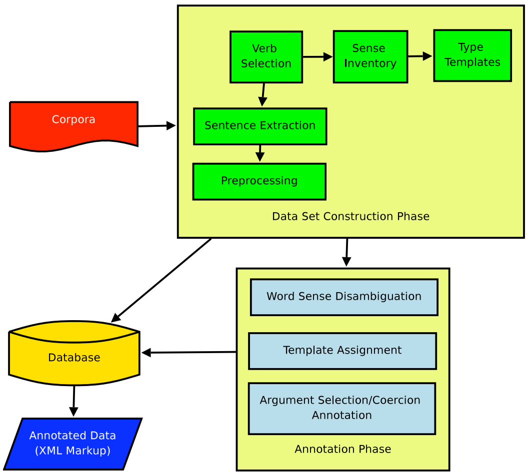

Figure 2: Corpus Development Architecture

4.1 Data Set Construction Phase: English

For the English data set, the data construction phase was combined with the annotation phase. The data for the task was created using the following steps:

1. The verbs were selected by examining the data from the BNC, using the Sketch Engine (Kilgarriff et al., 2004) as described in (Rumshisky and Batiukova, 2008). Verbs that consistently impose semantic typing on one of their arguments in at least one of their senses (strongly coercive verbs) were included into the final data set: *arrive (at)*, *cancel*, *deny*, *finish*, and *hear*.

2. Sense inventories were compiled for each verb, with the senses mapped to OntoNotes (Pradhan et al., 2007) whenever possible. For each sense, a set of type templates was compiled using a modification of the CPA technique (Hanks and Pustejovsky, 2005; Pustejovsky et al., 2004): every argument in the syntactic pattern associated with a given sense was assigned a type specification. Although a particular sense is often compatible with more than one semantic type for a given argument, this was never the case in our data set, where no disjoint types were tested. The coercive senses of the chosen verbs were associated with the following type templates:

   a. *Arrive (at)*, sense *reach a destination or goal*: HUMAN arrive at LOCATION
   b. *Cancel*, sense *call off*: HUMAN cancel EVENT
   c. *Deny*, sense *state or maintain that something is untrue*: HUMAN deny PROPOSITION
   d. *Finish*, sense *complete an activity*: HUMAN finish EVENT
   e. *Hear*, sense *perceive physical sound*: HUMAN hear SOUND

   We used a subset of semantic types from the Brandeis Shallow Ontology (BSO), which is a shallow hierarchy of types developed as a part of the CPA effort (Hanks, 2009; Pustejovsky et al., 2004; Rumshisky et al., 2006). Types were selected for their prevalence in manually identified selection context patterns developed for several hundred English verbs. That is, they capture common semantic distinctions associated with the selectional properties of many verbs. The types used for annotation were:

   ABSTRACT ENTITY, ANIMATE, ARTIFACT, ATTITUDE, DOCUMENT, DRINK, EMOTION, ENTITY, EVENT, FOOD, HUMAN, HUMAN GROUP, IDEA, INFORMATION, LOCATION, OBLIGATION, ORGANIZATION, PATH, PHYSICAL OBJECT, PROPERTY, PROPOSITION, RULE, SENSATION, SOUND, SUBSTANCE, TIME PERIOD, VEHICLE

   This set of types is purposefully shallow and non-hierarchical. For example, HUMAN is a subtype of both ANIMATE and PHYSICAL OBJECT, but annotators and system developers were instructed to choose the most relevant type (e.g., HUMAN) and to ignore inheritance.

3. A set of sentences was randomly extracted for each target verb from the BNC (Burnard, 1995). The extracted sentences were parsed automatically, and the sentences organized according to the grammatical relation the target verb was involved in. Sentences were excluded from the set if the target argument was expressed as anaphor, or was not present in the sentence. The semantic head for the target grammatical relation was identified in each case.

4. Word sense disambiguation of the target predicate was performed manually on each extracted sentence, matching the target against the sense inventory and the corresponding type templates as described above. The appropriate senses were then saved into the database along with the associated type template.

5. The sentences containing coercive senses of the target verbs were loaded into the Brandeis Annotation Tool (Verhagen, 2010). Annotators were presented with a list of sentences and asked to determine whether the argument in the specified grammatical relation to the target belongs to the type associated with that sense in the corresponding template. Disagreements were resolved by adjudication.

Coerion Type	Verb	Train	Test
EVENT→LOCATION	*arrive at*	38	37
ARTIFACT→EVENT	*cancel*	35	35
	finish	91	92
EVENT→PROPOSITION	*deny*	56	54
ARTIFACT→SOUND	*hear*	28	30
EVENT→SOUND	*hear*	24	26
DOCUMENT→EVENT	*finish*	39	40

Table 1: Coercions in the English data set

6. To guarantee robustness of the data, two additional steps were taken. First, only the six most recurrent coercion types were selected; these are given in table 1. Preference was given to cross-domain coercions, where the source and the target types are not related ontologically. Second, the distribution of selection and coercion instances were skewed to increase the number of coercions. The final English data set contains about 30% coercions.

7. Finally, the data set was randomly split in half into a training set and a test set. The training data has 1032 instances, 311 of which are coercions, and the test data has 1039 instances, 314 of which are coercions.

4.2 Data Set Construction Phase: Italian

In constructing the Italian data set, we adopted the same methodology used for the English data set, with the following differences:

1. The list of coercive verbs was selected by examining data from the ItWaC (Baroni and Kilgarriff, 2006) using the Sketch Engine (Kilgarriff et al., 2004):

   accusare 'accuse', *annunciare* 'announce', *arrivare* 'arrive', *ascoltare* 'listen', *avvisare* 'inform', *chiamare* 'call', *cominciare* 'begin', *completare* 'complete', *concludere* 'conclude', *contattare* 'contact', *divorare* 'devour', *echeggiare* 'echo', *finire* 'finish', *informare* 'inform', *interrompere* 'interrupt', *leggere* 'read', *raggiungere* 'reach', *recar(si)* 'go to', *rimbombare* 'resound', *sentire* 'hear', *udire* 'hear', *visitare* 'visit'.

2. The coercive senses of the chosen verbs were associated with type templates, some of which are listed listed below. Whenever possible, senses and type templates were adapted from the Italian Pattern Dictionary (Hanks and Jezek, 2007) and mapped to their SIMPLE equivalents (Lenci et al., 2000).

   a. *arrivare*, sense *reach a location*: HUMAN arriva [prep] LOCATION
   b. *cominciare*, sense *initiate an undertaking*: HUMAN comincia EVENT
   c. *completare*, sense *finish an activity*: HUMAN completa EVENT
   d. *udire*, sense *perceive a sound*: HUMAN ode SOUND
   e. *visitare*, sense *visit a place*: HUMAN visita LOCATION

   The following types were used to annotate the Italian dataset:

   ABSTRACT ENTITY, ANIMATE, ARTIFACT, ATTITUDE, CONTAINER, DOCUMENT, DRINK, EMOTION, ENTITY, EVENT, FOOD, HUMAN, HUMAN GROUP, IDEA, INFORMATION, LIQUID, LOCATION, ORGANIZATION, PHYSICAL OBJECT, PROPERTY, SENSATION, SOUND, TIME PERIOD, VEHICLE

   The annotators were provided with a set of definitions and examples of each type.

3. A set of sentences for each target verb was extracted and parsed from the *PAROLE sottoinsieme corpus* (Bindi et al., 2000). They were skimmed to ensure that the final data set contained a sufficient number of coercions, with proportionally more selections than coercions. Sentences were preselected to include instances representing one of the chosen senses.

4. In order to exclude instances that may have been wrongly selected, a judge performed word sense disambiguation of the target predicate in the extracted sentences.

5. Annotators were presented with a list of sentences and asked to determine the usual semantic type associated with the argument in the specified grammatical relation. Every sentence was annotated by two annotators and one judge, who resolved disagreements.

6. Some of the coercion types selected for Italian were:

   a. LOCATION → HUMAN (*accusare, annunciare*)
   b. ARTIFACT → HUMAN (*annunciare, avvisare*)
   c. EVENT → LOCATION (*arrivare, raggiungere*)
   d. ARTIFACT → EVENT (*cominciare, completare*)
   e. EVENT → DOCUMENT (*leggere, divorare*)
   f. HUMAN → DOCUMENT (*leggere, divorare*)
   g. EVENT → SOUND (*ascoltare, echeggiare*)
   h. ARTIFACT → SOUND (*ascoltare, echeggiare*)

7. The Italian training data contained 1466 instances, 381 of which are coercions; the test data had 1463 instances, with 384 coercions.

5 Data Format

The test and training data were provided in XML. The relation between the predicate (viewed as a function) and its argument were represented by composition link elements (`CompLink`), as shown below. The test data differed from the training data in the omission of `CompLink` elements.

In case of *coercion*, there is a mismatch between the source and the target types, and both types need to be identified; e.g., *The State Department repeatedly denied the attack*:

```
The State Department repeatedly
<SELECTOR sid="s1">denied</SELECTOR>
the <TARGET id="t1">attack</TARGET>.
<CompLink cid="cid1"
          compType="COERCION"
          selector_id="s1"
          relatedToTarget="t1"
          sourceType="EVENT"
          targetType="PROPOSITION"/>
```

When the compositional operation is *selection*, the source and target types must match; e.g., *The State Department repeatedly denied the statement*:

```
The State Department repeatedly
<SELECTOR sid="s2">denied</SELECTOR>
the <TARGET id="t2">statement</TARGET>.
<CompLink cid="cid2"
          compType="SELECTION"
          selector_id="s2"
          relatedToTarget="t2"
          sourceType="PROPOSITION"
          targetType="PROPOSITION"/>
```

6 Results & Analysis

We received only a single submission for the ASC task. The **UTDMet** system was an SVM-based system with features derived from two main sources: a PageRank-style algorithm over WordNet hypernyms used to define semantic classes, and statistics from a PropBank-style parse of some 8 million documents from the English Gigaword corpus. The results, shown in Table 2, were computed from confusion matrices constructed for each of four classification tasks for the 1039 link instances in the English test data: determination of argument selection or coercion, identification of the argument source type, identification of the argument target type, and the joint identification of the source/target type pair.

Clearly, the UTDMet system did quite well at this task. The one immediately noticeable outlier is the macro-averaged precision for the joint type, which reflects a small number of miscategorizations of rare types. For example, eliminating the single miscategorized ARTIFACT-LOCATION link in the submitted test data bumps this score up to a respectable 94%. This large discrepancy can explained by the lack of *any* coercions with those types in the gold-standard data.

	Prec.	Recall	Averaging
Selection *vs.* Coercion:	95	96	(macro)
	96	96	(micro)
Source Type:	96	96	(macro)
	96	96	(micro)
Target Type:	100	100	(both)
Joint Type:	86	95	(macro)
	96	96	(micro)

Table 2: Results for the UTDMet submission.

In the absence of any other submissions, it is difficult to provide a point of comparison for this performance. However, we can provide a baseline by taking each link to be a selection whose source and target types are the most common type (EVENT for the gold-standard English data). This yields micro-averaged precision scores of 69% for selection *vs.* coercion, 33% for source type identification, 37% for the target type identification, and 22% for the joint type.

The performance of the UTDMet system suggests that most of the type coercions were identifiable based largely on examination of lexical clues associated with selection contexts. This is in fact to be expected for the type coercions that were the focus of the English data set. It will be interesting to see how systems perform on the Italian data set and an expanded corpus for English and Italian, where more subtle and complex type exploitations and manipulations are at play. These will hopefully be explored in future competitions.

7 Conclusion

In this paper, we have described the Argument Selection and Coercion task for SemEval-2010. This task involves identifying the relation between a predicate and its argument as one that encodes the compositional history of the selection process. This allows us to distinguish surface forms that directly satisfy the selectional (type) requirements of a predicate from those that are coerced in context. We described some details of a specification language for selection, the annotation task using this specification to identify argument selection behavior, and the preparation of the data for the task. Finally, we analyzed the results of the task submissions.

References

M. Baroni and A. Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of European ACL.*

R. Bindi, P. Baroni, M. Monachini, and E. Gola. 2000. PAROLE-Sottoinsieme. *ILC-CNR Internal Report.*

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of LREC*, Genoa, Italy.

L. Burnard, 1995. *Users' Reference Guide, British National Corpus.* British National Corpus Consortium, Oxford, England.

Marcus Egg. 2005. *Flexible semantics for reinterpretation phenomena.* CSLI, Stanford.

P. Hanks and E. Jezek. 2007. Building Pattern Dictionaries with Corpus Analysis. In *International Colloquium on Possible Dictionaries*, Rome, June, 6-7. Oral Presentation.

P. Hanks and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée.*

P. Hanks. 2009. Corpus pattern analysis. CPA Project Page. Retrieved April 11, 2009, from http://nlp.fi.muni.cz/projekty/cpa/.

J. R. Hobbs, M. Stickel, and P. Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex, Lorient, France*, pages 105–116.

Karin Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon.* Phd dissertation, University of Pennsylvania, PA.

A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, et al. 2000. SIMPLE: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249.

K. Markert and M. Nissim. 2007. SemEval-2007 task 8: Metonymy resolution. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June. Association for Computational Linguistics.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31.

Geoffrey Nunberg. 1979. The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, 3:143–184.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

S. Pradhan, E. Hovy, MS Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *International Conference on Semantic Computing, 2007*, pages 517–526.

J. Pustejovsky, P. Hanks, and A. Rumshisky. 2004. Automated Induction of Sense in Context. In *COLING 2004, Geneva, Switzerland*, pages 924–931.

J. Pustejovsky, R. Knippen, J. Littman, and R. Sauri. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39(2):123–164.

J. Pustejovsky, A. Rumshisky, J. Moszkowicz, and O. Batiukova. 2009. GLML: Annotating argument selection and coercion. *IWCS-8: Eighth International Conference on Computational Semantics.*

J. Pustejovsky. 1991. The generative lexicon. *Computational Linguistics*, 17(4).

A. Rumshisky and O. Batiukova. 2008. Polysemy in verbs: systematic relations between senses and their effect on annotation. In *COLING Workshop on Human Judgement in Computational Linguistics (HJCL-2008)*, Manchester, England.

A. Rumshisky, P. Hanks, C. Havasi, and J. Pustejovsky. 2006. Constructing a corpus-based ontology using model bias. In *The 19th International FLAIRS Conference, FLAIRS 2006*, Melbourne Beach, Florida, USA.

J. Ruppenhofer, M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice.*

Carlos Subirats. 2004. FrameNet Español. Una red semántica de marcos conceptuales. In *VI International Congress of Hispanic Linguistics*, Leipzig.

Marc Verhagen. 2010. The Brandeis Annotation Tool. In *Language Resources and Evaluation Conference, LREC 2010*, Malta.